ED 090 304 TM 003 592

AUTHOR Kidder, Carole L.; Golub, Lester S.

TITLE Computer Application of a Syntactic Density Measure.

PUB DATE Apr 74

NOTE 12p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, Illinois, April, 1974)

EDRS PRICE MF-$0.75 HC-$1.50 PLUS POSTAGE

DESCRIPTORS Complexity Level; *Computer Programs; Educational Diagnosis; Elementary School Students; *Language Development; *Measurement Techniques; *Readability; *Syntax

ABSTRACT

Recent research in syntactic complexity has described and specified structural characteristics that distinguish levels of syntactic density. A computer program, written in Programmer Language I for the IBM 370 at Pennsylvania State University, has been developed to apply a syntactic density formula to samples of natural language and provide a single quantitative score. The program is being used to assess syntactic density as a factor of readability in samples of graded reading material and to measure stages in children's language development as evidenced in written language samples from children in elementary school. (Author)

Carole L. Kidder

Lock Haven State College

Lock Haven, Pennsylvania

Lester S. Golub

The Pennsylvania State University

University Park, Pennsylvania

COMPUTER APPLICATION OF A SYNTACTIC DENSITY MEASURE

It is empirically obvious that there are varying degrees of complexity in syntax in different levels of graded reading materials and in the oral and written language of children at different levels of development. Research in language education and in language development could be facilitated by using the computer to analyze and measure syntactic density (or syntactic complexity or syntactic maturity.)

One of the characteristics of children's language development is that with increasing maturity, children use more and more complex structures. Research has shown (Hunt, 1965, 1970; Loban, 1963, 1970; O'Donnell, Griffin, and Norris, 1967) that even after the pre-school period of rapid language acquisition, students of elementary and secondary school ages continue to develop abilities to manipulate language by employing more complicated syntactic structures. Much of the recent research in syntactic development (Hunt, 1970; Loban, 1970; Golub and Frederick, 1971) has been aimed at discovering, describing, and specifying those characteristics of syntax that distinguish degrees of complexity, maturity, or density of syntax. Out of this research have come some instruments that provide single, quantitative scores to distinguish between levels of syntactic density.

When any of these instruments is applied to language samples, hand tabulation is tedious, time-consuming, and subject to the inconsistencies of human error. Some of the instruments require that the analyst have some degree of sophistication in linguistic analysis. The cost, in time and training, of hand analysis inhibits research designs that require analysis of large samples of natural language. Significant research findings in many possible studies of language development would require sizable samples of text. Analysis of syntactic density could also be useful for comparison of stylistic characteristics of speakers or authors and for assessment of syntactic load as a factor of readability.

Since research (Golub, 1971) has identified and described specific syntactic features that indicate increased degrees of syntactic density, the next logical step seems to be to program the computer to apply the instruments to language samples for fast, efficient, and consistent results.

Chomsky repeatedly asserts that language performance cannot be equated with language competence. At best, measurement of performance can only give an indication of competence. Yet, realizing that we cannot discover everything about idealized language competence, we should not be prevented from learning about improved methods of measuring performance.

As the child learns to put words together in meaningful relationships he is developing a grammar that enables him to generate an increasing variety of unique sentences. Through his own processes of observation, classification,' hypothesis-making, and hypothesis-testing, he moves more and more toward the adult model of his language community. In the early stages of language development, surface structures and deep structures are relatively isometric, and utterance units are frequently kernel-like units. As conceptual ability, vocabulary, and relational abilities develop along with increasingly powerful rules for sentence formation and transformation, the map from surface structure to deep structure becomes more complicated. What might have been several communication units before becomes a single unit, more fully packed with meanings which are manipulated by more complicated syntactic structures. The syntactic density increases.

To allow quantitative comparisons of syntactic complexity in samples of language, an agreed-upon basic unit of comparison has had to be found. Early research (LaBrant, 1933) found the sentence too subjective a measure. Hunt has defined the T-unit word length, a main clause with all of its subordinate clauses, as a more reliable measure. Subsequent research (Hunt, 1970; O'Donnell, 1967) has substantiated the reliability of the T-unit measure and found that with increasing maturity T-unit length tends to increase.

Within the last fifteen years research has contributed valuable information about the types of structures and amount of their use in the oral and written language of children at various ages. Some of these studies were conducted before the advent of transformational grammar. Some have dealt transformationally with limited age ranges. Among the measures that Hunt found to be signficantly related to increased syntactic maturity were: T-unit length, subordinate clause length, and reductions to less than a predicate. O'Donnell, Griffin, and Norris (1967) found that T-unit length, number of sentence-combining transformations, and deletion transformations contributed substantially to structural complexity when oral and written samples of language from children in grades 3, 5, and 7 were analyzed.

Through a series of studies of children's oral and written discourse, Golub has developed a Syntactic Density instrument that tabulates the occurrences of specific linguistic structures that correlate with teachers' judgments of writing samples. In an early stage of the study, a sixty-three linguistic variables were listed. Multivariate analysis isolated the ten variables that most highly correlated with teachers' high ratings. Canonical correlation assigned a relative weight to each variable according to the degree of its contribution to "syntactic density." The resulting Tabulation Sheet for a SYNTACTIC DENSITY SCORE provides for a calculation. When the variables are counted and weighted, the products are added. The total is divided by the number of T-units in the sample to arrive at a single syntactic density score. The variables included in Golub's formula reflect structures that have been identified in linguistic theory as being complex structures. Measures of mean main clause length and mean subordinate clause length are combined with measures of these other types of complexities.

Insert Figure 1 about here

Golub's formula for measuring syntactic density has been selected as the instrument to be programmed for the computer. It incorporates the measures of T-unit length and subordinate clause length that Hunt and others have found useful. It also reflects complex verb expansions, use of some advanced structures of time, and reductions or embeddings that take the form of prepositional phrases. Hand-tabulation by Golub's formula is rather time consuming and requires some training for the rater.

A program for use on an IBM 370 computer has been written in PL/1 by Carole Kidder to apply the formula to samples of natural language.

Encoding Conventions for Data

Text to be analyzed by the computer program must be keypunched, or typed on a Remote Job Entry, in blank-delimited form in columns 1 to 72. This means that each word and syntactic punctuation mark must be preceded and followed by at least one blank. Lexical punctuation, such as in hyphenated words or in abbreviations, is not separated from its associated character string by blanks. Multiple blanks are ignored. Quotation marks surrounding

Figure 1

SYNTACTIC DENSITY SCORE

Tabulation Sheet

Variable Number	Variable Description	Variable Loading	Frequency	VLXF
	Total number of words			
	Total number of T-units			
1.	Words/T-unit	.95	X ________	________
2.	Subordinate clauses/T-unit	.90	X ________	________
3.	Main clause word length (mean)	.20	X ________	________
4.	Subordinate clause word length (mean)	.50	X ________	________
5.	Number of Modals (will, shall, can, may, must, would.....)	.65	X ________	________
6.	Number of Be and Have forms in the auxiliary	.40	X ________	________
7.	Number of Prepositional Phrases	.75	X ________	________
8.	Number of Possessive nouns and pronouns	.70	X ________	________
9.	Number of Adverbs of Time (when, then, once, while...)	.60	X ________	________
10.	Number of gerunds, participles, and absolute phrases (unbound modifiers)	.85	X ________	________

Total ________

SDS
S.D. Score (Total/No. of T-units) ________

Grade Level Conversion ________

Grade Level Conversion Table:

SDS	.5	1.3	2.1	2.9	3.7	4.5	5.3	6.1	6.9	7.7	8.5	9.3	10.1	10.9
Grade Level	1	2	3	4	5	6	7	8	9	10	11	12	13	14

conversation may be omitted. At the end of each paragraph, the final mark of punctuation must be doubled. To separate samples, at the end of each sample three dollar signs ($$$) must appear in columns 1 to 3.

The first job step in the program is an Indexer, that picks off the individual words of all the text to be analyzed and puts them in a structure that indexes each one according to linear number with respect to the entire text, word in sentence, sentence number, paragraph number, author number, and sample group number.[1] The output record of the Indexer is stored on a temporary systems disc to be fed into the next job step, the Analyzer.

Analyzer

The Analyzer step of the job begins by giving the computer some reference lists, or dictionaries, and some decision-making capacity. Stop lists are initialized into the program so that the computer can reference and cross reference coordinating conjunctions, subordinating conjunctions, relative pronouns, modals, forms of "have" and "be", prepositions, possessive pronouns, and adverbs of time. A six-by-twelve transition matrix, designed by Paul Schuepp of Pennsylvania State University, houses the decision power. The accompanying diagram (Figure 2) displays a conceptualization of the Transition Matrix. Figure 3 is a summary of the Transition Matrix routines that control the program.

Insert Figure 2 and Figure 3 about here

The indexed text is brought into core one sentence at a time. Each word or syntactic punctuation mark is analyzed by the matrix. Punctuation, stop lists, and their interrelationships are examined until a decision can be

[1] This indexer is a modified version of the index feature of John Smith's RATSCAN (1972).

Figure 2

CONCEPTUALIZATION OF THE TRANSITION MATRIX

CLASS STATE		. ! ; 0	COMMA , 1	QUES-TION ? 2	COLON : 3	POSS-ESIVE & - 's - s' 4	-ING -ED -EN 5	COORD INATE CONJ. 6	SUBOR DINATE CONJ. &R.P. 7	HAVE -BE FORMS 8	MODAL IN AUX. 9	PREP-OSI-TION 10	TIME AD-VERBS 11	EVERY THING ELSE 12
T-UNIT	0	End -T 0	GNW 1	Q 0	GNW 1	POSS 0	VERB -AL 0	FOR 0	SC 4	HAVE OR BE 2	MOD-AL 0	PREP 0	TIME 0	GNW 0
FOLLOWING COMMA PATHS	1	ERR 1 0	ERR 1 0	ERR 1 0	ERR 1 0	POSS 2	VERB -AL 2	CS 0	SC 4	HAVE OR BE 2	MOD-AL 2	PREP 2	TIME 2	GNW 2
POSSIBLE ITEMS IN A SERIES 1	2	END -T 0	SERI-ES 0	Q 0	SERI -ES 0	POSS 3	VERB -AL 3	FOR 0	SC 4	HAVE OR BE 3	MOD-AL 3	PREP 3	TIME 3	GNW 3
POSSIBLE ITEMS IN A SERIES 2	3	END -T 0	SERI -ES 0	Q 0	SERI -ES 0	POSS 0	VERB -AL 0	FOR 0	SC 4	HAVE OR BE 0	MOD-AL 0	PREP 0	TIME 0	GNW 0
POSSIBLE SUBORD-CLAUSE 1	6	END -T 0	GNW 1	Q 0	SERI -ES 0	POSS 5	VERB -AL 5	FOR 5	SKIP 4	HAVE OR BE 5	MOD-AL 5	PREP 5	TIME 5	GNW 5
POSSIBLE SUBORD-CLAUSE 2	5	END -T 0	GNW 1	Q 0	SERI -ES 0	MARK SUB 6	MARK SUB 6	MARK SUB 6	SKIP 7 5	MARK SUB 6	MARK SUB 6	MARK SUB 6	MARK SUB 6	MARK SUB 6
DEFINITE SUBORD-CLAUSE	6	END -T 0	GNW 0	Q 0	SERI -ES 0	POSS 6	VERB -AL 6	FOR 6	GNW 6	HAVE OR BE 6	MOD-AL 6	PREP 6	TIME 6	GNW 6

Figure 3

SYNTACTIC DENSITY ROUTINES SUMMARY FROM TRANSITION MATRIX

1. END-T: End of a T-unit is encountered. Calculate main clause word count. Initialize other variables, and read in the next sentence. Then go to GNW.

2. GNW: Get next word in sentence to determine class it is in. Then determine state. Branch to that entry in matrix and perform that routine.

3. SC: Check for subordinate conjunctions longer than one word (in order that, so that, provided that) and for those that cross-reference with prepositions.

4. SERIES: Flag items in series. If a coordinating conjunction appears next, do not let it flag a compound sentence.

5. CS: If items-in-a-series has not been flagged, a compound sentence has probably been encountered. Increment T-unit count.

6. FOR: If the coordinating conjunction is for, since it did not follow a comma, for must be a preposition here. Increment preposition count.

7. MARK-SUB: Since three words have followed the subordinate conjunction and no punctuation has been encountered, mark as a subordinate clause and increment subordinate clause word count by 3. Increment subordinate conuunction count by 1.

8. Q: A question mark is encountered. Check to see if the first word in the sentence is a relative pronoun. If so, cancel subordinate clause markers.

9. VERBAL: Check for words ending in -ing, -ed, or -en, that have more than 6 characters and that do not have a form of have or be within the preceding 3 words. If all 3 conditions prevail, increment verbal count.

10. POSSES: Check to see if the word is one of the possessive pronouns or ends in -'s or -s'. If one of these conditions is met, increment possessives count.

11. HAVE-BE: Check to see if the word is a form of haveor be. If so, increment ve-be count.

12. MODALS: Check the reference list to see if the word is a modal. If so, increment modal count.

13. TIME: Check to see if the word is on the list of adverbs of time. If so, increment adverbs of time count.

14. ERR 1: Print out the sentence for evaluation.

    A possible punctuation error has been encountered.

15. ERR 2: Print out the sentence for evaluation.

    A possible undefined transition matrix entry may be causing an error.

made about what routine should be called. The routines process other program algorithms and implement tabulation, flagging, or computation.

For most variables and for all computations, the machine scoring has been found to be more accurate than hand scoring. Many of the decisions to be made by the machine are quite deterministic. Counts of possessive pronouns, modals, and words, for instance, can be definitely and easily decided. For more complicated decisions, program algorithms check series of conditions to be met before a decision is made. A few of the decisions are probabilistic. Occasionally one of the probabilistic decisions might be discovered to be erroneous, but repeated analyses reveal that the program is consistent and predominantly accurate. The printout shows the text being analyzed and a tabulation sheet for each sample that lists the frequencies and subscores on each of the linguistic variables and gives the computed Syntactic Density Score. A Grade Level Conversion Table is also displayed on each Tabulation Sheet.

To compare the machine scoring with hand scoring, twelve 200-word samples of graded reading material were scored by a trained rater and carefully checked by a second rater. The same samples were then scored by the computer. The Pearson Product-Moment Correlation of the "hand" and "machine" analysis was .96, with the machine scores running consistently slightly higher than the hand scores. Golub's original formula calls for a count of forms of *have* or *be* used in the auxiliary position. The computer counts all occurrences of forms of *have* or *be*. This difference in hand tabulation and computer tabulation accounts for some of the slightly higher score when machine and hand scores are compared.

Sixty 200-word samples of graded reading material (ten samples at each grade level from second to seventh grade) were analyzed by the program to determine whether there is a significant difference in syntactic density in materials designed for different grade levels. The Syntactic Density increases at each grade level. The differences between grade levels were statistically significant at yearly intervals in two of the five intervals ($\underline{p} \leqslant .05$). When differences were examined by two-year intervals, statistical significance was found at every interval ($\underline{p} \leqslant .05$).

An analysis of written language of children in first through sixth grade is now in progress to test the discriminatory power and range of the instrument. The computer program is also being used in a study of the effects of alternate types of learning experiences on children's written compositions, in a project to coordinate reading levels of selected and directed reading materials for a Pennsylvania school district, and in preparing some materials for testing and research in content areas.

The Syntactic Density computer program should be a useful instrument for further experimental investigations, for assessing readability levels, for development of materials, for diagnostic purposes, and for stylistic analysis.

References

Botel, Morton and Granowsky, Alvin. "A Formula for Measuring Syntactic Complexity: A Directional Effort," Elementary English, 49, April, 1972, pp. 513-516.

Golub, Lester S. and Fredrick, Wayne C. Linguistic Structures and Deviations in Children's Written Sentences. Technical Report from the Wisconsin Research and Development Center for Cognitive Learning. The University of Wisconsin, No. 152, 1970.

Golub, Lester S. and Fredrick, Wayne C. Linguistic Structures in the Discourse of Fourth and Sixth Graders. Technical Report from the Wisconsin Research and Development Center for Cognitive Learning. The University of Wisconsin, No. 166, July, 1971.

Hunt, Kellogg. Grammatical Structures Written at Three Grade Levels. National Council of Teachers of English Research Report No. 3. Champaign, Ill.: NCTE, 1965.

Hunt, Kellogg. Syntactic Maturity in Schoolchildren and Adults. Monographs of the Society for Reserach in Child Development, Vol. 35, No. 1, February, 1970.

Loban, Walter, The Language of Elementary School Children. Champaign, Ill.: National Council of Teachers of English, 1963.

Loban, Walter. Stages, Velocity, and Prediction of Language Development Kindergarten through Grade Twelve. Final Report, Project No. 7-0061, U.S. Department of Health, Education, and Welfare, 1970.

McCaig, R. A. "How Not to Analyze the Syntax of Children: A Critique and a Proposal." Elementary English, May, 1970, pp. 612-618.

O'Donnell, Roy C., Griffin, William J., and Norris, Raymond. Syntax of Kindergarten and Elementary School Children: A Transformational Analysis. National Council of Teachers of English Report No. 8. Champaign, Ill.: NCTE, 1967.

Smith, John B. "RATS: A Middle-Level Text Utility System." Computers and the Humanities, Vol. 6, May, 1972, pp. 277-283.

Strickland, Ruth G. The Language of Elementary School Children: Its Relationship to the Language of Reading Textbooks and the Quality of Reading of Selected Children. Bulletin of the School of Education, Indiana University, Vol. 38, No. 4, Bloomington, Indiana, July, 1962.

Templin, Mildred C. Certain Language Skills in Children, Their Development and Interrelationships. Minneapolis: The University of Minnesota Press, 1957.